AI Influence: Mechanisms, Amplifiers, and Consequences

Anonymous authors
Paper under double-blind review

Abstract

AI influence refers to AI's impact on the knowledge and values of individuals by acting as producers, mediators, and receivers of information. As a result, it impacts our collective processes of creating and spreading knowledge, forming beliefs, and reaching consensus. We argue that there are mechanisms of inconspicuous influence in AI development and deployment pipelines, which, when amplified by societal dynamics, could lead to dangerous outcomes that we may reverse by early interventions. We detail those mechanisms, amplifiers, and potential long-term consequences.

1 Introduction

1.1 Overview of AI Influence

AI influence refers to the impact of AI technologies on the knowledge and values of individuals, whether as a producer (e.g., LLM output), mediator (e.g., recommender system), or receiver (e.g., preference learning from human feedback) of information. As a result, it further impacts our collective processes of creating and spreading knowledge, forming beliefs, and reaching consensus. Different from "AI's impact in general", which concerns AI's broad societal impacts encompassing economic, legal, social, and environmental dimensions, "AI Influence" specifically zeroes in on the epistemic and axiological dimensions of this impact. It is concerned with how AI alters how humans know (epistemology) and what humans value (axiology), rather than just the direct outcomes of AI deployment.

We propose "AI influence" to unify scattered research efforts. Empirical research on AI influence is ongoing but scattered. Those efforts are either clustered around specific affected subjects — Wikipedia (Wagner & Jiang, 2025), Stack Exchange community (Burtch et al., 2024), open-source community (Yeverechyahu et al., 2024), scientific publication and peer review (Liang et al., 2024a;b), political campaigns and elections (Hackenburg & Margetts, 2024a; Potter et al., 2024) — or carved up along discipline boundaries like machine learning, cognitive science, education, human-AI interaction, and epistemology, with little cross-disciplinary discourse taking place.

AI influence is not necessarily a harm. Despite that AI influence on human epistemology poses serious concerns, it is too early to conclude that AI influence is, on net, a bad thing. Humans are bound by cognitive limitations, and it's likely that AI may expand our cognitive capacity and improve our collective deliberation.

1.2 Our Contributions

Proposing AI influence as a distinctive research field AI influence concerns the pervasive, subtle, and long-term ways AI reconfigures human internal states — in particular, the cognitive processes of acquiring and using knowledge, forming beliefs, and making judgments. This is distinct from AI safety or AI ethics research, where researchers address direct harms imposed on marginalized groups or engineering failures of AI systems (Gross, 2023; Hendrycks & Mazeika, 2022).

Introducing a three-level framework for AI influence We decompose the landscape of AI influence into three dimensions: *mechanisms*, the basic channels through which AI influences human epistemology; *amplifiers*, external factors that significantly enlarge such influence; and *consequences*, societal hazards that

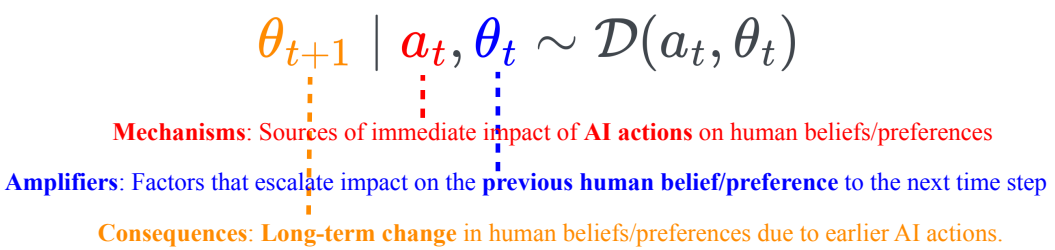

Figure 1: Formalizing the distinction between mechanisms, amplifiers, and consequences.

the amplified influence have led to or may soon lead to. These are *orthogonal* dimensions often connected by fully interconnected relationships, which allows us to focus on each of them individually while also touching on their connections.

Summarizing methodologies for studying AI influence We summarize all the methods that are used so far to study AI influence in Table 2. We also reason from first principles what could be used to study the influence from technologies on human truth-seeking and morality development process. We illuminate the gap and discuss future research directions.

	Language Models (Supervised)	Language Models (Reinforced)	Recommender Systems	Knowledge-Based Systems
Context Space $\mathcal{C}$	Natural-language prompts		User-item pairs	Assertions/attributes
Action Space $\mathcal{A}$	Natural-language responses		Scores/rankings	(Truth) Values
Distribution $\mathcal{D}_\mathcal{C}$	Training-time distribution of contexts			Expert input coverage
Policy Space Π	Parameterized deep neural networks			Deterministic assignments
Loss Function $\mathcal{L}_\theta$	Cross-entropy loss against human response	Plackett-Luce disagreement with human preference	Disagreement with human scoring/ranking	Abidance with expert-sourced constraints (0/1)
Regularizer Ω	Weight decay	KL regularization	Social regularization	Logical consistency constraints (0/1)

Table 1: How four example types of AI system training/development can be mapped onto the common formalism. Each column represents one simplified, canonical example of that type of system, with only distinctive features shown.

2 Absence of AI Influence in Existing Paradigms

A large class of intelligent systems are designed to interact with humans in ways that alter human cognition. Current examples of such systems include:

- **Language Models and Language Agents.** Language models undergo pretraining on massive amounts of human-generated data (Devlin et al., 2019), which equip them for interaction with human users on information-seeking tasks. More importantly, the alignment training on these models (Ji et al.) through reinforcement learning from human feedback (Bai et al., 2022), direct preference optimization (Rafailov et al., 2023), or personalization methods (Chen et al., 2024), makes them produce user-preferred behaviors in explicitly interactive setups.

- **Recommender and Ranking Systems.** These systems function as information gatekeepers that filter vast item spaces to present users with prioritized content. Architectures in this domain have evolved from matrix factorization techniques (Koren et al., 2009) to deep candidate generation and ranking models (Covington et al., 2016). While typically framed as predictors of static user preferences based on historical interaction data, modern iterations frequently employ reinforcement learning objectives to maximize long-term engagement rewards (Chen et al., 2019). This optimization process treats the user's internal state as a dynamic component of the environment. Consequently, the system actively shapes user preferences by determining the exposure distribution of information and altering the choice architecture available to the agent.
- **Knowledge-Based and Decision Support Systems.** Knowledge-based systems often utilize structured data representations, including knowledge graphs and ontologies (Wang et al., 2017), or hybrid neuro-symbolic architectures that integrate deep learning with rule-based constraints (Belle, 2020). Unlike open-ended generation, these tools operate within strictly defined action spaces to output risk scores, diagnostic suggestions, or factual retrievals. The epistemic influence here manifests through automation bias and anchoring (Bansal et al., 2021). By presenting calculated probabilities or retrieved facts with high system confidence, these tools establish authoritative baselines that shift the human decision boundary regardless of the underlying ground truth.

We focus on this class of AI systems in the rest of this paper and avoid distinctions between them, as their commonalities induce the mechanisms, amplifiers, and consequences that we will introduce. In this section, we first dissect the currently mainstream formalisms that guide the training of these systems, and then specify the categories of missing elements.

2.1 Common Formalism for Learning From Humans

One common pattern emerge upon examining the mainstream formalisms in the development and training of all major types of AI systems that we consider. This high-level pattern involves a space $\mathcal{C}$ of *contexts*, a distribution $\mathcal{D}_\mathcal{C}$ over contexts, a space $\mathcal{A}$ of *actions*, and a space $\Pi \subseteq \Delta[\mathcal{A}]^\mathcal{C}$ of *policies* mapping any context to a distribution over actions. The task is then to solve the optimization problem

$$\text{minimize } \mathrm{E}_{c\sim\mathcal{D}_\mathcal{C}}\left[\mathcal{L}_\theta(c,\pi(c))\right] + \Omega(\pi), \quad \text{s.t. } \pi \in \Pi,$$

where $\mathcal{L}_\theta(\cdot,\cdot)$ is a (parameterized) loss defined against human annotation, and $\Omega(\cdot)$ a regularization term.

Table 1 explains how the training or development of the aforementioned types of systems — language models (Devlin et al., 2019; Bai et al., 2022), recommender systems (Resnick & Varian, 1997), and knowledge-based systems (Akerkar & Sajja, 2009) — can be mapped to such a formalism. It also implies what the parameter θ stands for, i.e., a human *belief state*, describing the beliefs, preferences, etc. of the human in the training loop.

There is the further dimension of *time*. The common formalism above is typically used to accomodate time in one of two ways.

1. *Assuming Full Stationarity.* When the loss function and the context distribution is stationary and the policy is memory-less, the interaction at each time step is simply a replay of the previous time steps. As such, the temporal setup can be directly reduced to the one-off setup. This is the case in, for example, language model training (Bai et al., 2022; Hadfield-Menell et al., 2016).
2. *Assuming Stationarity of Loss Function.* In formalisms such as Markov decision processes, the context (state) depends on the interaction history, and the policy is memory-ful and maps a pair of context *and history* to an action distribution. However, they still assume a stationary loss (reward) function, i.e., $\theta_t = \theta$, for some fixed reward parameter θ. This is the case in, for example, most recommender systems trained with reinforcement learning (Afsar et al., 2022).

In reality, however, neither approaches fully suffice, as the implicit assumption of a stationary loss function ($\theta_t = \theta$) fails. It fails due to the natural shift and development of human beliefs/preferences, and, importantly,

due to the presence of *AI influence*. In other words, the loss function (parameterized by θ_t) at the t-th time step is often a function of the previous actions taken by the AI system.

Note that θ_t can either be the state of a single human source (e.g., a user convinced of conspiracy theories who now gives conspiracy theory-aligned preference feedback), or that of a human collective (e.g., a company's hiring distribution being biased by a discriminatory hiring decision support system, which feeds back into the system's training data).

In the following subsections, we examine this missing influence in the training paradigms, from three precisely defined angles (Figure 1).

2.2 Mechanisms: Sources of Immediate Influence

Figure 1 shows how the loss function $\mathcal{L}_{\theta_{t+1}}$ (decided by the preferences and beliefs of the humans that the system learns from) at time step $t+1$ depends on both the previous time step's human preferences and beliefs (θ_t) and the action taken by the AI system (a_t).

In Section 3, we will qualitatively explore the direct *mechanisms* of AI influence, i.e., the causal pathways through which the AI system's action a_t directly and immediately impacts the human belief state θ_{t+1} at the next time step. Ranging from persuasion (Durmus et al., 2024) to reliance (Nirman et al., 2024), such impact is empirically well-established through human studies (Jakesch et al., 2023; Glickman & Sharot, 2024b) and has been theoretically studied in a reinforcement learning setup (Carroll et al., 2024).

2.3 Amplifiers: Temporal Factors that Escalate Impact

In contrast to mechanisms, we define *amplifiers* as the factors that lead to the escalation of AI impact on the human belief state from the previous time step (θ_t) to the next (θ_{t+1}), i.e., the causal dependence of θ_{t+1} on θ_t that results in compounding errors.

Such factors can be endogenous, such as human confirmation bias (Oeberst & Imhoff, 2023), or exogenous, such as echo chambers formed by human-AI interaction (Glickman & Sharot, 2024b; Sharma et al., 2024), or institutional factors that entrench biased consensus (Lawrence et al., 2001; Bisson et al., 2021). We will discuss these factors in more detail in Section 4.

2.4 Consequences: Impact on Belief States Over Time

When both mechanisms of influence and amplifiers of influence are present, they tend to imply that the impact of AI actions on the human belief state θ_t compounds over time, eventually leading to large and potentially irreversible changes. Again, such changes can either be at the individual levels (beliefs, preferences) or at the collective level (norms, culture, collective knowledge).

In Section 5, we qualitatively characterize some key potential consequences of amplified AI influence. Some of them have already seen strong empirical evidence, while others are currently speculative predictions about long-term outcomes.

3 Mechanisms

In this section, we cover specific mechanisms through which AI systems play a direct and immediate role in influencing human epistemics and morality, at an individual level and societal level. By *mechanisms*, we refer to either technical limitations of AI systems or new ways through which humans interact with AI that directly causes change in human epistemics.

Here, we emphasize that the AI systems change how information is originated, disseminated, propagated, and received by humans or AI systems. The scope extends beyond that of algorithmic biases.

Table 2: Related research classified by methodology and topic. Empty cells indicate the lack of known works. Due to the relative scarcity of qualitative studies, we include them as a single category while using a fine-grained partition for quantitative studies.

		Qualitative Research	Formal Models	Simulations	Descriptive Analysis	Observational Causal Inference	RCTs
Mechanisms	Digital Reliance	Gerlich (2025); Hirvonen et al. (2024); Glickman & Sharot (2024a); Brandtzaeg et al. (2024); Kulveit et al. (2025); Kirk et al. (2025)			Gerlich (2025); Burtch et al. (2024); Nirman et al. (2024); Thompson et al. (2024); Wagner & Jiang (2025); Dillion et al. (2025)	Burtch et al. (2024)	Kruegel et al. (2025)
	Distinct AI Biases	Köbis et al. (2021); Brandtzaeg et al. (2024)	Taori & Hashimoto (2022)	Haroon et al. (2023)	Adilazuarda et al. (2024); Barman et al. (2024); Lamparth et al. (2024); Ryan et al. (2024); Agarwal et al. (2025); Weng et al. (2025); Wang et al. (2025)	Brown et al. (2022); Haroon et al. (2023); Yakura et al. (2024)	Glickman & Sharot (2024b); Fisher et al. (2024); Danry et al. (2024); Costello et al. (2024a); Kidd & Birhane (2023b); Kruegel et al. (2025); Leib et al. (2021); Piccardi et al. (2024); Potter et al. (2024)
	Direct Persuasion	Brandtzaeg et al. (2025); Hirvonen et al. (2024)	Collins et al. (2025); Dean et al. (2024)	Ferraro et al. (2024)	Fisher et al. (2024)	Hosseinmardi et al. (2024)	Argyle et al. (2025); Hackenburg & Margetts (2024a); White et al. (2024); Goel et al. (2025); Costello et al. (2024a); Durmus et al. (2024); Matz et al. (2024); Danry et al. (2025)
	Attention Reallocation	Schuster & Lazar (2025)				Su et al. (2016)	Mendler-Dünner et al. (2024b); Haupt et al. (2023); Hosseinmardi et al. (2024)
Amplifiers	Human-AI Dual Influence	Brady et al. (2023); Brady & Crockett (2024); Collins et al. (2024); Lazar et al. (2024); Li & Yin (2024); Pedreschi et al. (2025)	Lin et al. (2024); Ferbach et al. (2024); Collins et al. (2025); Qiu et al. (2025); Krueger et al. (2020)	Wang et al. (2024); Brinkmann et al. (2022); Ferraro et al. (2024); Collins et al. (2025); Qiu et al. (2025); Mansoury et al. (2020); Perra & Rocha (2019); Dean et al. (2024)	Li et al. (2023); Liang et al. (2024a)	Qiu et al. (2025)	Glickman & Sharot (2024b); Brinkmann et al. (2022); Chan et al. (2024); Pataranutaporn et al. (2023); Haupt et al. (2023); Hosseinmardi et al. (2024); Lu et al. (2024); Pappalardo et al. (2024); Sharma et al. (2024)
	Trust				Dillion et al. (2025); Araujo et al. (2020); Helberger et al. (2020)		Narayanan et al. (2023); Pataranutaporn et al. (2023); Reis et al. (2024); Osborne & Bailey (2025)
	Institutional Path Dependence	Kulveit et al. (2025); Simon & Isaza-Ibarra (2023); Aoki (2024); Gruetzemacher et al. (2024); Lazar & Manuali (2024); Matz et al. (2024); Ovadya et al. (2024); Leibo et al. (2025)	Jarrett et al. (2025)	Zhang et al. (2025); Jarrett et al. (2025)			Potter et al. (2024)
	Socio-Economic Matthew Effect	Capraro et al. (2024)			Wang et al. (2025)	Su et al. (2016)	
Consequences	Lock-in of Human Errors		Collins et al. (2025); Qiu et al. (2025); Lin et al. (2024)	Qiu et al. (2025); Collins et al. (2025); Wang et al. (2024); Mansoury et al. (2020); Perra & Rocha (2019)		Qiu et al. (2025)	Chan et al. (2024); Costello et al. (2024a); Haupt et al. (2023); Kubin & Sikorski (2021)
	Lock-in of AI Biases	Brandtzaeg et al. (2024); Kulveit et al. (2025); Köbis et al. (2021)	Taori & Hashimoto (2022)	Piao et al. (2025)	Adilazuarda et al. (2024); Barman et al. (2024); Lamparth et al. (2024); Ryan et al. (2024)		Glickman & Sharot (2024b); Fisher et al. (2024); Danry et al. (2024); Costello et al. (2024a); Kidd & Birhane (2023b); Kruegel et al. (2025); Leib et al. (2021); Piccardi et al. (2024); Potter et al. (2024)
	Value Capture	Nguyen (2024b)					
	Knowledge Collapse	Brandtzaeg et al. (2024); Glickman & Sharot (2024a); Koskinen (2024); Wihbey (2024)	De Marzo et al. (2024)	Peterson (2024); Bossens et al. (2024)	Burtch et al. (2024); Dohmatob et al. (2024); Thompson et al. (2024); Li et al. (2023); Liang et al. (2024a); Si et al. (2024); Wagner & Jiang (2025); Wu et al. (2024)	Anderson et al. (2024)	Doshi & Hauser (2023); Padmakumar & He (2023); Sharma et al. (2024)
	Epistemic Stratification	Kay et al. (2024)			Agarwal et al. (2025); Wang et al. (2025)		

3.1 AI Introduces Distinct Biases into Collective Knowledge

Although AI systems are trained on data generated by humans, they do acquire distinctive biases from humans (Glickman & Sharot, 2024b; Kahneman et al., 2021). Specifically, there are the following reasons that introduce distinctive AI biases:

- *Learning systems like LLMs are struggling with long-tail knowledge.* As a primary example of learning-based intelligent systems, the question-answering accuracy of LLMs correlate strongly with how many times questions and answers co-occur in the training dataset (Kandpal et al., 2023; Das et al., 2024).
- *Architectures create unique AI biases.* Architectural biases often stem from technical limitations, as opposed to biases in datasets that can be more readily resolved by more training or more data. One notable example is the bias of Convolutional Neural Networks (CNNs) towards texture (Geirhos et al., 2018). Tokenization, the strategy LLM employs to split words into subwords, introduces biases unique to AI, such as downgrading arithmetic performances (Singh & Strouse, 2024), mishandling grammatical structures, and biases in handling rare words (Phan et al., 2024).

Through training and deploying AI systems that acquire distinct biases, we risk introducing new biases into the collective knowledge-making process, such as publication, journalism, scientific research, etc. Such AI biases might be persistent or even amplified because of digital reliance or feedback loops, as we will discuss in the following two subsections.

3.2 Cognitive Offloading, Cognitive Enhancements, and Digital Reliance

AI can enhance human cognitive performance, which can take place either directly by providing advice and implementable solutions (Senior et al., 2020; Fawzi et al., 2022) or indirectly by revealing novel cognitive strategies and problem-solving approaches (Shin et al., 2023). Cognitive offloading is the term commonly used to describe such activities, namely, physical actions (such as preparing a grocery list) to reduce cognitive demands required (Risko & Gilbert, 2016). Research shows that humans are willing to offload attention-demanding tasks to AI systems (Wahn et al., 2023). AI systems are also used to improve human cognitive performance. For example, a study that examines the performance of Go players (Shin et al., 2023) reveals that the performance of Go players improved after being exposed to AlphaGo moves, possibly as a result of learning novel non-human strategies from AlphaGo. Consistent results come from a study examining human problem-solving in a navigation task (Brinkmann et al., 2022). In this study, participants navigated through complex networks. Each path was associated with rewards (earning points) or penalties (losing points). Before performing the task, participants were exposed to solutions generated by the AI or by humans. The results demonstrated enhanced performance (accumulation of higher rewards) among players learning from AI, mainly due to the exposure to counterintuitive but optimal strategies generated by the AI. For example, the AI better identified than humans paths that initially appeared suboptimal but ultimately yielded better outcomes.

On the other hand, those cognitive offloading and enhancement activities enabled by AI may lead to digital reliance. Research demonstrates that reliance on digital tools, and in particular AI, alters different cognitive processes such as memory, critical thinking, and problem-solving. For example, Sparrow et al. (2011) showed that when information is accessible through search engines, individuals prioritize remembering where to find this information rather than retaining it. This pattern extends to modern AI systems as well. Gerlich (2025) found that cognitive offloading to AI tools correlates with reduced critical thinking engagement, particularly among younger users who exhibit higher dependency. Consistent with these empirical findings, Zhai et al. (2024) conducted a systematic review revealing that over-reliance on AI dialogue systems impairs critical thinking and decision-making by fostering cognitive shortcuts. Together, these studies suggest that in some cases, delegating cognitive tasks to AI systems may deteriorate fundamental cognitive and thinking capabilities.

In the context of this paper, digital reliance makes space for bias amplification, as we will discuss in the following subsections.

3.3 AI Persuasion Directly Reshapes Human Beliefs

As AI systems become integral to how humans access and evaluate information, they exert increasing influence over the processes of belief and opinion formation. Recent studies demonstrate that conversational AI can measurably shape political attitudes (Hackenburg & Margetts, 2024b; Fisher et al., 2025) and alter support for electoral candidates (Argyle et al., 2025). While persuasive capacity generally scales with model size, post-training methods and prompting strategies can yield even larger effects (Durmus et al., 2023; Hackenburg & Margetts, 2024b). Consequently, smaller models with targeted fine-tuning can achieve persuasive capabilities comparable to frontier systems, rendering influence tools broadly accessible. Techniques that maximize persuasive effectiveness, such as information-dense rhetoric, are associated with systematic reductions in factual accuracy, indicating a potential trade-off between persuasive power and epistemic reliability (Hackenburg & Margetts, 2024b).

Although such persuasive capabilities raise concerns about manipulation for financial or political gain, they may also be directed toward prosocial ends. Targeted human–AI dialogues have been shown to increase effective charitable giving beyond either static AI messages or human appeals (White et al., 2024). Similarly, conversational interventions with AI can durably reduce conspiratorial thinking, with effects persisting for up to two months (Costello et al., 2024b), and can decrease confidence in false beliefs (Goel et al., 2025). At the individual level, the epistemic influence of AI can be seen as dual-use: it can amplify both epistemic risk and epistemic improvement, depending on the underlying objectives.

3.4 AI Reallocates Human Attention

One of the major functionalities of AI systems is that they reorganize and redistribute information available to us, as search engines (including LLM-based ones) and RecSys-based social media do. In the previous subsection, we cover new mechanisms through which AI biases affect human judgements, while in this case, AI influences what we see and think by selecting what information gets presented to us and receives our attention. This may have a strong agenda-setting effect on our thinking (Mendler-Dünner et al., 2024a).

We elaborate on the problem of attention allocation and the resulting segmentation of users. For sophisticated users of AI technologies, it is possible for generative models to be hugely creative, adding to intellectual diversity (Meincke et al., 2024). But such possibilities require careful technique and strategy, from few-shot prompting to chain-of-thought reasoning to iterative strategies in general. For the vast majority of the model-using public, who may not understand what the models are and do, and have little ability to execute prompt engineering strategies, usage may be largely passive and simplistic. Models will therefore tend to provide answers and content to the majority of users that conform to mainstream, modal patterns — the most likely next token, the probabilistic best answer or idea. This, in fact, is their central tendency and what they are designed to do. Using the models in a simplistic auto-complete or recommendation engine-style is likely to direct human attention to mainstream ideas and trends that are featured prominently on the open web (where the model pre-training has taken place), and not necessarily to more diverse, challenging, obscure, or marginal ideas or viewpoints.

4 Amplifiers

Mechanisms enumerated in Section 3 explain the forces that AI systems exert on human cognition and epistemology. Those forces tend to be subtle and may not pose extreme risks on their own.

Meanwhile, in this section, we introduce a range of *amplifiers* that are *external* to AI systems and may significantly increase AI influence (usually social factors), to the degree of posing systemic risks described in Section 5.

4.1 Human-AI Dual Influence Creates Feedback Loops

The influence between AI and humans is not one-directional. Humans' preferences can be influenced by the content generated by AI systems, while AI systems are trained to align with human preferences as well (e.g., Reinforcement Learning with Human Feedback (Ziegler et al., 2019)). Such a feedback loop between humans

and AI is similar to the feedback loop between content users and content creators in recommender systems, where users' tastes are shaped by the content they consume and creators produce content to fit users' tastes (Jiang et al., 2019; Lin et al., 2024).

Although human-AI dual influence might help to improve the alignment between humans and AI, it could also bring potential harm. For example, when humans or AI have initial biases or errors regarding a certain topic, such biases and errors can be circulated and amplified in human-AI interactions. There has been extensive research on human-to-AI and AI-to-human influence, but it was not until very recently that research showed human-AI interaction may further exacerbate this influence mechanism: biased AI systems can affect human beliefs, rendering humans more biased compared to the initial state, due to the amplification of bias by AI systems and assigned trust by humans in AI judgments (Glickman & Sharot, 2024b;a).

AI bias is an established research field (Mayson, 2018). In this paper, however, we argue that digital reliance on AI and feedback loops established in human-AI interactions legitimize larger concerns over this topic. Not only because bias affects the accuracy of medical decisions (Challen et al., 2019) or racial fairness (Salinas et al., 2023), which are by themselves important problems, but also because those biases are permanently introduced into epistemic processes and alter our worldviews (Vicente & Matute, 2023).

4.2 Trust Amplifies AI Influence

Do higher levels of trust in AI correlate with increased AI influence? Recent research provides evidence supporting this claim. For example, Vicente & Matute (2023) demonstrated that higher trust in AI systems in medical diagnostic tasks led participants to adopt more of AI's biased recommendations, and even carry these into subsequent tasks. Similarly, it was found that self-reported trust in AI systems was associated with the persuasiveness of deceptive AI classifications; interestingly, trust was not associated with the effect of improved AI-generated truthful explanations in the case of news headlines (Danry et al., 2024), although results to the contrary were found in a credit loan decision-making setup (Sunny, 2025).

Current evidence suggests that human trust in AI is highly sensitive to context and culture. While in many contexts, people prefer AI advice over humans' (Araujo et al., 2020; Logg et al., 2019), in high-stakes contexts (such as medicine or other life-threatening cases), people assign trust to humans more than AI systems (Reis et al., 2024). Additionally, Globig et al. (2024) found that trust in AI varies significantly across cultures (Globig et al., 2024). Individuals in Eastern countries (e.g., India, Indonesia) exhibit greater trust and optimism towards AI compared to their Western counterparts (e.g., U.S., Germany), who tend to be more skeptical and cautious (Globig et al., 2024).

4.3 Institutional Path Dependence

Institutional path dependence refers to the tendency of organizations and systems to make decisions and adopt practices based on past trajectories, often locking in early patterns of behavior (Page et al., 2006). Epistemic frameworks through which institutions understand and address issues can be influenced by AI, an influence that can be hard to remove given the self-reinforcing nature of institutions (Arthur, 2018).

For instance, widespread AI application in the education sector may plant deep-rooted AI influence in children (Xu & Ouyang, 2022), AI advisors and analytics may bias governmental decision-making processes toward specific data-driven perspectives (Castelnovo & Sorrentino, 2021), AI-influenced public opinion can reinforce or challenge institutional norms (Panait & Ashraf, 2021), and early critical attitudes toward AI-generated art and writing have led to the enactment of institutional policies against the use of language models (Takagi, 2023; Kreitmeir & Raschky, 2023). Once these AI-mediated epistemic influences take root, their self-reinforcing nature may make it difficult to shift away from initial decisions, even in light of new evidence or changing contexts.

The self-reinforcing nature of the institutional path dependence problem will be particularly difficult to mitigate, given recursion (Peterson, 2024). Once embedded narratives take hold and the climate of human opinion gets expressed at scale on social media and the web, AI models themselves will subsequently be trained on this new data containing AI influence. This "data coil" means path dependence becomes difficult to resist or reverse (Beer, 2022).

4.4 Socio-Economic Matthew Effect

Advanced AI systems threaten to dramatically amplify existing socio-economic inequalities through what we term the "AI Matthew Effect", whereby initial advantages in AI access and capability compound exponentially over time.

Specifically, AI Matthew Effect occurs when groups initially receiving more benefit from AI (*e.g.*, the wealthy, speakers of majority languages, those living in developed nations, those with access to GPUs, those working in fields where training data is more abundant) receive cascading benefits, and vice versa. An example is when biases against minority languages in LLMs shrink their user base who speak minority languages, which could further reinforce biases against minority languages due to under-representation.

This dynamic could manifest through several interconnected mechanisms:

Productivity amplification: AI systems act as force multipliers for human productivity, with their effectiveness scaling in proportion to the user's existing capabilities and resources. High-skilled knowledge workers with access to state-of-the-art AI tools can leverage them to augment their expertise, potentially increasing their productivity by orders of magnitude. Meanwhile, workers in lower-skilled positions may find their jobs automated or devalued, creating a widening productivity gap.

Capital concentration: Organizations with early access to powerful AI systems can optimize operations, reduce costs, and capture market share more effectively than competitors. This advantage creates a self-reinforcing cycle where increased profits enable further AI investment and development, leading to market concentration.

5 Consequences

Influence mechanisms (Section 3), whose effects are magnified by amplifiers (Section 4), may lead to long-term consequences that are associated with large-scale hazards.

Long-term consequences are hard to clearly demonstrate in advance, but some have nonetheless manifested in empirical studies. Here we make a non-exhaustive list of these potential consequences.

AI systems that are trained on human data contains human errors and biases (Mayson, 2018; Binz & Schulz, 2023; Yax et al., 2024). Direct and indirect interactions with those models can circulate those biases back to humans (Morewedge et al., 2023; Valyaeva, 2024). Furthermore, those human errors and biases can be amplified via human-AI interactions because humans may assign more trust in AI output than average humans (Logg et al., 2019). These psychological traits of humans and the training methods of learning systems (e.g. LLMs) raise concerns that those human errors and biases might be permanently preserved, amplified, and even locked into human society over the long run. The term "lock-in" refers to cases where values, beliefs, knowledge, and practices are introduced into human society, last for a long time, spread widely, assume a dominant memetic position in a population, are institutionalized (therefore hard to remove), and cause damage (Hendrycks & Mazeika, 2022).

5.1 Lock-in of AI Biases

AI bias has been well documented and studied — not only in the realms of fairness and equality (Bolukbasi et al., 2016; Caliskan et al., 2017), but also on broadly construed biases in cultural and factual domains (Santurkar et al., 2023).

However, a consequential effect has been largely overlooked: when humans interact with these biased systems, they internalize the systems' amplified bias and become more biased than they initially were (Glickman & Sharot, 2024b; Vicente & Matute, 2023). This bias amplification feedback loop relies on two key characteristics of AI systems: First, AI systems provide a higher signal-to-noise ratio compared to humans, consistently producing less variable outputs than human judgments (Kahneman et al., 2021). Second, in many domains, humans perceive AI systems as more capable and accurate than other humans (Logg et al., 2019), making them more receptive to AI influence or uncritically adopting AI biases. For instance, clinicians inherit

AI biases even after AI systems are removed (Vicente & Matute, 2023). These characteristics create a dynamic where even small initial biases can be rapidly adopted and magnified through human-AI interactions. Furthermore, this effect raises particular concerns for children, who have more malleable knowledge structures and may be more susceptible to AI's influence than adults (Kidd & Birhane, 2023a), raising the concern that such AI biases would be locked-in over generations.

5.2 Goodhart's Law and Value Capture

Human objectives are often operationalized into quantifiable metrics — for instance, research quality being quantified as citation counts, and idea quality being quantified as the number of retweets. In economics, Goodhart's law, *"when a measure becomes a target, it ceases to be a good measure,"* states that optimizing for a quantifiable proxy initially leads to improvement in the true objective, but beyond a certain point, such optimization often leads to (potentially catastrophic) degradation in the true objective (Goodhart, 1984).

An instance of Goodhart's law in the case of human values, *value capture*, happens when one mistakes quantified proxies for their much richer terminal values, and exclusively optimizes for the former instead, thereby losing the ability of personal deliberation on their values (Nguyen, 2024a).

AI has already been used in such quantification of objectives, for example, in social media (Anandhan et al., 2018); other similar uses of AI has also been proposed, including as arbiters for resolving human disagreement (Tessler et al., 2024) and human representatives for collective decision-making (Zhang et al., 2024). In all such cases, human actors may be incentivised, or are already incentivised (Lüders et al., 2022; Wolf et al., 2017), to optimize for the AI-defined objectives. If such optimization becomes the dominant concern of human participants — which is plausible given that AI products are often designed to be game-like and addictive (De et al., 2025) — value capture may steer people's values and objectives away from an ideal deliberative choice.

5.3 Knowledge Collapse

Knowledge collapse (Peterson, 2024) is defined as *the progressive narrowing over time of the set of information available to humans, along with a concomitant narrowing in the perceived availability and utility of different sets of information.* It is hypothesized to manifest as a "mode collapse" of collective knowledge in the human community, where long-tail information is lost while mainstream information is strengthened.

Peterson (2024) mainly focuses on unrepresentative data, lack of in-depth exploration during LLM inference, and algorithmic limitations of next-token prediction as the potential causes of knowledge collapse. Peterson (2024) argues that by making mainstream information more readily available, learning systems like LLMs shift attention away from long-tail information.

In addition to these concerns, we note that other mechanisms outlined in this paper, including, for example, dual influence (Lin et al., 2024), can similarly contribute to knowledge collapse. From a mechanistic angle, knowledge collapse and lock-in share many commonalities, most especially the reinforcement of existing popular ideas and the suppression of marginal ones.

5.4 Epistemic Stratification

Epistemic stratification is the unequal distribution of access to knowledge, resources, and cognitive tools across individuals or groups, leading to disparities in their ability to acquire, evaluate, and generate knowledge (Silva Filho et al., 2023).

AI may contribute to epistemic stratification by amplifying existing disparities, such as through unequal access to advanced AI tools, biased algorithmic recommendations that reinforce echo chambers, the prioritization of information access for privileged demographics (Kay et al., 2024), or the increasingly centralized control over AI development (Brynjolfsson & Ng, 2023).

6 Caveats and Counterarguments

6.1 AI Systems Have Negligible Influence on Human Cognition

Empirical evidence does not provide a holistic picture of AI's impact on human cognition. It is true that humans are becoming more reliant on AI systems for their tasks, but it is unclear whether having AI systems to process those tasks for humans would necessarily degenerate or enhance those cognitive capabilities of humans. At least, it is still unclear whether humans' navigation skills are compromised because of using GPS tools (Fricker, 2021; Jadallah et al., 2017).

Two questions are instrumental to understanding AI systems' impact on human cognition. For one, does digital reliance influence human cognitive skills that are directly replaced by corresponding AI capabilities (Teschke et al., 2013)? For instance, does the use of GPS hurt human navigation skills (Fricker, 2021)? The same question could be asked about other digital tools and human skills, such as calculators and arithmetic skills, machine translation tools and second-language acquisition. For the other, do the replaced domain-specific human skills undermine more general human cognitive capabilities? For example, do the undermined arithmetic skills hurt human general mathematical reasoning and problem-solving abilities (Geary et al., 2015; Hurst & Cordes, 2018)?

Without sufficient empirical evidence on how human cognition might be altered in the presence of new tools, especially AI systems, it is hard to firmly hold our position. Hence, an alternative view is, AI systems may have a negligible impact on human cognitive capabilities over the long term. One reason is that we do not understand the relationship between low-level domain-specific skills and high-level general capabilities. Replacing the former by AI may have little negative impact on the latter, in which case adequate tool use may actually enhance cognitive capabilities (Teschke et al., 2013).

6.2 Highly Parameterized AI Systems Are Less Biased and Error-Prone Than Humans Are

AI systems are biased (Jadallah et al., 2017) and error-prone (Zhou et al., 2024), as research has revealed, but so are humans. Besides those inductive biases that are introduced by specific architectures and training methods (Geirhos et al., 2018; Singh & Strouse, 2024), AI systems acquire their biases from training datasets and, by extension, from humans. Highly parameterized AI systems such as LLMs are less biased and error-prone than conventional machine learning models as they are more expressive, and techniques such as RAG help to consult external sources for truth validation (Gao et al., 2023). Meanwhile, it is also likely that state-of-arts AI systems may become even less biased and error-prone than average humans are. From the point of view of collective truth-seeking (such as conducting scientific research and collective deliberation), AI systems functioning as "shadow authors" to individual humans can be positive.

That being said, err should be on the side of being cautious. It is likely that AI biases, errors, and hallucinations become more elusive before they are removed (Zhou et al., 2024). Once they are hard to find for average users, commercial developers are much motivated to address those problems, creating persistent and even amplified biases and errors (Ren et al., 2024), which are precisely what we warn in this paper.

6.3 AI's Epistemic Influence Can Be Positive

In Section 5, we have detailed AI's long-term impact on human knowledge and values. Notably, they seem overwhelmingly negative. It is not our intention to present negative views only, but we are likely biased and limited in our perspectives. We want to raise attention on AI's epistemic influence and avoid the cascading effects over the long term, but we also want to acknowledge that we are far from having a holistic picture.

It is entirely likely the issues we have raised here can be addressed over time and people can become wise in using those tools. For instance, users, especially students and researchers, may acquire a critical lens of AI-generated content. Under the name of "AI literacy", students are taught to use, understand, and evaluate AI systems critically (Casal-Otero et al., 2023). Sufficient critical thinking skills, paired with AI systems' increasing reach and capabilities, may cultivate a generation of more informed learners and citizens who are more capable of participating in collective truth-seeking and deliberative processes.

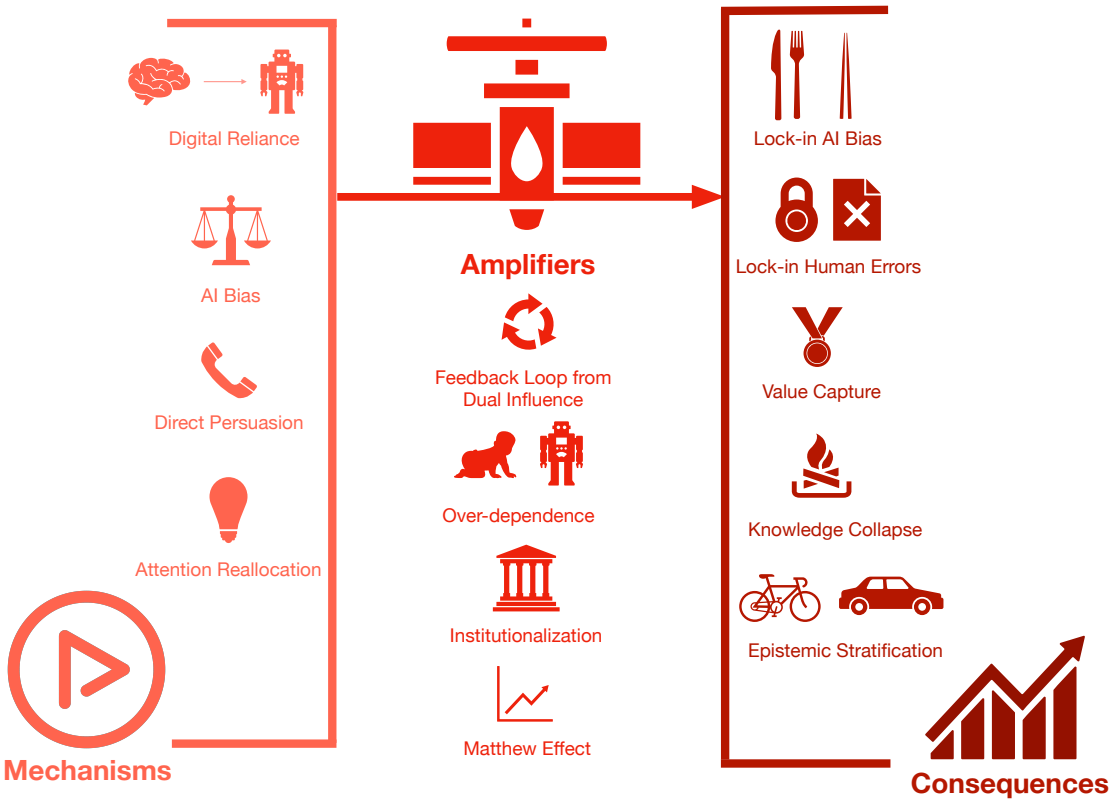

Figure 2: Mechanisms, amplifiers, and consequences of AI influence.

7 Conclusion

AI exerts systematic influence over the beliefs and values in individuals and society. We have outlined the mechanisms that enable such influence, the amplifiers that magnify the influence, and the potential consequences it may entail.

The eventual aim of AI influence research is to enable the responsible management of AI influence over human cognition, knowledge, and values, reaping its benefits while avoiding the harms. Accomplishing such an aim requires coordination between communities of interdisciplinary methodologies and perspectives, including AI safety and AI ethics communities, machine learning and human-computer interaction communities, social science communities, and, importantly, industry actors.

Broader Impact Statement

Recognizing AI influence is a necessary precondition for managing it, and in this respect, we aim to promote societal interest by raising awareness on the issue. Since mid 2025, reports of epistemic and psychological harms from deployed AI systems — extreme examples of which include AI-driven psychosis (TREYGER et al., 2025), with milder examples of influence and reliance being exponentially more prevalent (Phang et al., 2025) — have become increasingly prevalent. Given the trend of increasingly wide and immersive deployment of AI systems, it is likely that such epistemic and psychological impact will expand by orders of magnitude in the near future. We hope that increased awareness on this class of problems can foster the development of technical and governance solutions for the management of AI influence.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards Measuring and Modeling "Culture" in LLMs: A Survey. *arXiv*, 2024. doi: 10.48550/arxiv.2403.15412.

M Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. AI suggestions homogenize writing toward western styles and diminish cultural nuances. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2025. doi: 10.1145/3706598.3713564.

Rajendra Akerkar and Priti Sajja. *Knowledge-based systems*. Jones & Bartlett Publishers, 2009.

Anitha Anandhan, Liyana Shuib, Maizatul Akmar Ismail, and Ghulam Mujtaba. Social media recommender systems: review and open research issues. *IEEE Access*, 6:15608–15628, 2018.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization Effects of Large Language Models on Human Creative Ideation. *Creativity and Cognition*, pp. 413–425, 2024. doi: 10.1145/3635636.3656204.

Goshi Aoki. Large language models in politics and democracy: A comprehensive survey. *arXiv preprint arXiv:2412.04498*, 2024.

Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI & society*, 35(3):611–623, 2020.

Lisa P. Argyle, Ethan C. Busby, Joshua R. Gubler, Alex Lyman, Justin Olcott, Jackson Pond, and David Wingate. Testing theories of political persuasion using AI. *Proceedings of the National Academy of Sciences*, 122(18):e2412815122, May 2025. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2412815122. URL `https://pnas.org/doi/10.1073/pnas.2412815122`.

W Brian Arthur. Self-reinforcing mechanisms in economics. In *The economy as an evolving complex system*, pp. 9–31. CRC Press, 2018.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–16, 2021.

Kristian González Barman, Simon Lohse, and Henk de Regt. Reinforcement Learning from Human Feedback: Whose Culture, Whose Values, Whose Perspectives? *arXiv*, 2024.

David Beer. The problem of researching a recursive society: Algorithms, data coils and the looping of the social. *Big Data & Society*, 9(2):20539517221104997, 2022.

Vaishak Belle. Symbolic logic meets machine learning: A brief survey in infinite domains. In *International conference on scalable uncertainty management*, pp. 3–16. Springer, 2020.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

Linda F Bisson, Philip H Kass, Kyaw Tha Paw U, and Laura Grindstaff. Assessing institutionalized bias. In *Uprooting Bias in the Academy: Lessons from the Field*, pp. 61–80. Springer, 2021.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

David M Bossens, Shanshan Feng, and Yew-Soon Ong. The Digital Ecosystem of Beliefs: does evolution favour AI over humans? *arXiv*, 2024. doi: 10.48550/arxiv.2412.14500.

William J. Brady and M. J. Crockett. Norm Psychology in the Digital Age: How Social Media Shapes the Cultural Evolution of Normativity. *Perspectives on Psychological Science*, 19(1):62–64, 2024. ISSN 1745-6916. doi: 10.1177/17456916231187395.

William J. Brady, Joshua Conrad Jackson, Björn Lindström, and M.J. Crockett. Algorithm-mediated social learning in online social networks. *Trends in Cognitive Sciences*, 27(10):947–960, 2023. ISSN 1364-6613. doi: 10.1016/j.tics.2023.06.008.

Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. Understanding model power in social AI. *AI & SOCIETY*, pp. 1–11, 2024. ISSN 0951-5666. doi: 10.1007/s00146-024-02053-4.

Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. Understanding model power in social AI. *AI & SOCIETY*, 40(4):2839–2849, April 2025. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-024-02053-4. URL `https://link.springer.com/10.1007/s00146-024-02053-4`.

L. Brinkmann, D. Gezerli, K. V. Kleist, T. F. Mller, I. Rahwan, and N. Pescetelli. Hybrid social learning in human-algorithm cultural transmission. *Philosophical Transactions of the Royal Society A*, 380(2227): 20200426, 2022. ISSN 1364-503X. doi: 10.1098/rsta.2020.0426.

Megan A Brown, James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. Echo chambers, rabbit holes, and algorithmic bias: How youtube recommends content to real users. *Available at SSRN 4114905*, 2022.

Erik Brynjolfsson and Andrew Ng. Big ai can centralize decision-making and power, and that'sa problem. *Missing links in AI governance*, pp. 65, 2023.

Gordon Burtch, Dokyun Lee, and Zhichen Chen. The consequences of generative AI for online knowledge communities. *Scientific Reports*, 14(1):10413, 2024. doi: 10.1038/s41598-024-61221-0.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Valerio Capraro, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-François Bonnefon, Pablo Brañas-Garza, Luigi Butera, Karen M Douglas, Jim A C Everett, Gerd Gigerenzer, Christine Greenhow, Daniel A Hashimoto, Julianne Holt-Lunstad, Jolanda Jetten, Simon Johnson, Werner H Kunz, Chiara Longoni, Pete Lunn, Simone Natale, Stefanie Paluch, Iyad Rahwan, Neil Selwyn, Vivek Singh, Siddharth Suri, Jennifer Sutcliffe, Joe Tomlinson, Sander van der Linden, Paul A M Van Lange, Friederike Wall, Jay J Van Bavel, and Riccardo Viale. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, 3(6):pgae191, 2024. doi: 10.1093/pnasnexus/pgae191.

Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. Ai alignment with changing and influenceable reward functions. *arXiv preprint arXiv:2405.17713*, 2024.

Lorena Casal-Otero, Alejandro Catala, Carmen Fernández-Morante, Maria Taboada, Beatriz Cebreiro, and Senén Barro. Ai literacy in k-12: a systematic literature review. *International Journal of STEM Education*, 10(1):29, 2023.

Walter Castelnovo and Maddalena Sorrentino. The nodality disconnect of data-driven government. *Administration & Society*, 53(9):1418–1442, 2021.

Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ quality & safety*, 28(3):231–237, 2019.

Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F Loftus. Conversational AI Powered by Large Language Models Amplifies False Memories in Witness Interviews. *arXiv*, 2024.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42, 2024.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 456–464, 2019.

Katherine M. Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E. Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, Adrian Weller, Joshua B. Tenenbaum, and Thomas L. Griffiths. Building machines that learn and think with people. *Nature Human Behaviour*, 8(10):1851–1863, 2024. doi: 10.1038/s41562-024-01991-9.

Katherine M Collins, Umang Bhatt, and Ilia Sucholutsky. Revisiting rogers' paradox in the context of human-AI interaction. *arXiv*, 2025. doi: 10.48550/arxiv.2501.10476.

Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714):eadq1814, 2024a. ISSN 0036-8075. doi: 10.1126/science.adq1814.

Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714):eadq1814, September 2024b. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adq1814. URL `https://www.science.org/doi/10.1126/science.adq1814`.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pp. 191–198, 2016.

Valdemar Danry, Pat Pataranutaporn, Matthew Groh, Ziv Epstein, and Pattie Maes. Deceptive AI systems that give explanations are more convincing than honest AI systems and can amplify belief in misinformation. *arXiv*, 2024.

Valdemar Danry, Pat Pataranutaporn, Matthew Groh, and Ziv Epstein. Deceptive Explanations by Large Language Models Lead People to Change their Beliefs About Misinformation More Often than Honest Explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–31, Yokohama Japan, April 2025. ACM. ISBN 9798400713941. doi: 10.1145/3706598.3713408. URL `https://dl.acm.org/doi/10.1145/3706598.3713408`.

Debarati Das, Karin De Langis, Anna Martin-Boyle, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Anugrah Hayati, Risako Owan, Bin Hu, Ritik Parkar, et al. Under the surface: Tracking the artifactuality of llm-generated data. *arXiv preprint arXiv:2401.14698*, 2024.

Debasmita De, Mazen El Jamal, Eda Aydemir, and Anika Khera. Social media algorithms and teen addiction: Neurophysiological impact and ethical considerations. *Cureus*, 17(1), 2025.

Giordano De Marzo, Pietro Gravino, and Vittorio Loreto. Recommender systems may enhance the discovery of novelties. *Journal of Physics: Complexity*, 5(4):045008, 2024.

Sarah Dean, Evan Dong, Meena Jagadeesan, and Liu Leqi. Accounting for ai and users shaping one another: The role of mathematical models. *arXiv preprint arXiv:2404.12366*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Danica Dillion, Debanjan Mondal, Niket Tandon, and Kurt Gray. Ai language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, 15(1):4084, 2025.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A Tale of Tails: Model Collapse as a Change of Scaling Laws. *arXiv*, 2024. doi: 10.48550/arxiv.2402.07043.

Anil R Doshi and Oliver P Hauser. Generative artificial intelligence enhances creativity but reduces the diversity of novel content. *arXiv*, 2023. doi: 10.48550/arxiv.2312.00506.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. *arXiv*, 2023. doi: 10.48550/arxiv.2306.16388.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models, 2024. URL `https://www.anthropic.com/news/measuring-model-persuasiveness`.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences. *arXiv*, 2024. doi: 10.48550/arxiv.2407.09499.

Antonino Ferraro, Antonio Galli, Valerio La Gatta, Marco Postiglione, Gian Marco Orlando, Diego Russo, Giuseppe Riccio, Antonio Romano, and Vincenzo Moscato. Agent-Based Modelling Meets Generative AI in Social Network Simulations. *arXiv*, 2024. doi: 10.48550/arxiv.2411.16031.

Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. Biased AI can Influence Political Decision-Making. *arXiv*, 2024.

Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W. Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. Biased AI can Influence Political Decision-Making, June 2025. URL `http://arxiv.org/abs/2410.06415`. arXiv:2410.06415 [cs].

Elizabeth Fricker. Should we worry about silicone chip technology de-skilling us? *Royal Institute of Philosophy Supplements*, 89:131–152, 2021.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

David C Geary, Mary K Hoard, Lara Nugent, and Jeffrey N Rouder. Individual differences in algebraic cognition: Relation to the approximate number and semantic memory systems. *Journal of experimental child psychology*, 140:211–227, 2015.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Michael Gerlich. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1):6, 2025. doi: 10.3390/soc15010006.

Moshe Glickman and Tali Sharot. AI-induced hyper-learning in humans. *Current Opinion in Psychology*, 60: 101900, 2024a. ISSN 2352-250X. doi: 10.1016/j.copsyc.2024.101900.

Moshe Glickman and Tali Sharot. How human–ai feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, pp. 1–15, 2024b.

Laura K Globig, Rachel Xu, Steve Rathje, and Jay J Van Bavel. Perceived (mis) alignment in generative artificial intelligence varies across cultures. *Preprint. DOI*, 10, 2024.

Natasha Goel, Thomas Bergeron, Blake Lee-Whiting, Danielle Bohonos, Md Mujahedul Islam, Sarah Lachance, Sonja Savolainen, Clareta Treger, and Eric Merkley. Artificial Influence: Comparing the Effects of AI and Human Source Cues in Reducing Certainty in False Beliefs. 2025. URL https://files.osf.io/v1/resources/2vh4k_v1/providers/osfstorage/6729096b29cfa10289f5d5c8?action=download&direct&version=2. Publisher: OSF.

Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice: The UK experience*, pp. 91–121. Springer, 1984.

Nicole Gross. What chatgpt tells us about gender: a cautionary tale about performativity and gender biases in ai. *Social Sciences*, 12(8):435, 2023.

Ross Gruetzemacher, Shahar Avin, James Fox, and Alexander K Saeri. Strategic Insights from Simulation Gaming of AI Race Dynamics. *arXiv*, 2024.

Kobi Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2403116121, 2024a.

Kobi Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2403116121, June 2024b. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2403116121. URL https://pnas.org/doi/10.1073/pnas.2403116121.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Muhammad Haroon, Magdalena Wojcieszak, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, and Zubair Shafiq. Auditing youtube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the national academy of sciences*, 120(50):e2213020120, 2023.

Andreas Haupt, Mihaela Curmei, François-Marie de Jouvencel, Marc Faddoul, Benjamin Recht, and Dylan Hadfield-Menell. The long-term effects of personalization: Evidence from youtube. 2023.

Natali Helberger, Theo Araujo, and Claes H. de Vreese. Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review*, 39:105456, 2020. ISSN 0267-3649. doi: 10.1016/j.clsr.2020.105456.

Dan Hendrycks and Mantas Mazeika. X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*, 2022.

Noora Hirvonen, Ville Jylhä, Yucong Lao, and Stefan Larsson. Artificial intelligence in the information ecosystem: Affordances for everyday information seeking. *Journal of the Association for Information Science and Technology*, 75(10):1152–1165, 2024. ISSN 2330-1635. doi: 10.1002/asi.24860.

Homa Hosseinmardi, Amir Ghasemian, Miguel Rivera-Lanas, Manoel Horta Ribeiro, Robert West, and Duncan J. Watts. Causally estimating the effect of YouTube's recommender system using counterfactual bots. *Proceedings of the National Academy of Sciences*, 121(8):e2313377121, 2024. ISSN 0027-8424. doi: 10.1073/pnas.2313377121.

Michelle Hurst and Sara Cordes. A systematic investigation of the link between rational number processing and algebra ability. *British Journal of Psychology*, 109(1):99–117, 2018.

May Jadallah, Alycia M Hund, Jonathan Thayn, Joel Garth Studebaker, Zachary J Roman, and Elizabeth Kirby. Integrating geospatial technologies in fifth-grade curriculum: Impact on spatial ability and map-analysis skills. *Journal of Geography*, 116(4):139–151, 2017.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-Writing with Opinionated Language Models Affects Users' Views. *arXiv*, 2023. doi: 10.48550/arxiv.2302.00560.

Daniel Jarrett, Miruna Pîslar, Michiel A Bakker, Michael Henry Tessler, Raphael Köster, Jan Balaguer, Romuald Elie, Christopher Summerfield, and Andrea Tacchetti. Language agents as digital representatives in collective decision-making. *arXiv*, 2025. doi: 10.48550/arxiv.2502.09369.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, et al. Ai alignment: A contemporary survey. *ACM Computing Surveys*.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate Feedback Loops in Recommender Systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 383–390, Honolulu HI USA, January 2019. ACM. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3314288. URL `https://dl.acm.org/doi/10.1145/3306618.3314288`.

Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. *Noise: A flaw in human judgment*. Hachette UK, 2021.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pp. 15696–15707. PMLR, 2023.

Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed. Epistemic Injustice in Generative AI. *arXiv*, 2024. doi: 10.48550/arxiv.2408.11441.

Celeste Kidd and Abeba Birhane. How ai can distort human beliefs. *Science*, 380(6651):1222–1223, 2023a.

Celeste Kidd and Abeba Birhane. How AI can distort human beliefs. *Science*, 380(6651):1222–1223, 2023b. ISSN 0036-8075. doi: 10.1126/science.adi0248.

Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A Hale. Why human-ai relationships need socioaffective alignment. *arXiv preprint arXiv:2502.02528*, 2025.

Nils Köbis, Jean-François Bonnefon, and Iyad Rahwan. Bad machines corrupt good morals. *Nature Human Behaviour*, 5(6):679–685, 2021. doi: 10.1038/s41562-021-01128-2.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Inkeri Koskinen. We Have No Satisfactory Social Epistemology of AI-Based Science. *Social Epistemology*, 38 (4):458–475, 2024. ISSN 0269-1728. doi: 10.1080/02691728.2023.2286253.

David H Kreitmeir and Paul A Raschky. The unintended consequences of censoring digital technology–evidence from italy's chatgpt ban. *arXiv preprint arXiv:2304.09339*, 2023.

Sebastian Kruegel, Andreas Ostermaier, and Matthias Uhl. ChatGPT's advice drives moral judgments with or without justification. *arXiv*, 2025. doi: 10.48550/arxiv.2501.01897.

David Krueger, Tegan Maharaj, and Jan Leike. Hidden Incentives for Auto-Induced Distributional Shift. *arXiv*, 2020.

Emily Kubin and Christian von Sikorski. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3):188–206, 2021. ISSN 2380-8985. doi: 10.1080/23808985.2021.1976070.

Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental AI development. *arXiv*, 2025. doi: 10.48550/arxiv.2501.16946.

Max Lamparth, Anthony Corso, Jacob Ganz, Oriana Skylar Mastro, Jacquelyn Schneider, and Harold Trinkunas. Human vs. Machine: Behavioral Differences Between Expert Humans and Language Models in Wargame Simulations. *arXiv*, 2024. doi: 10.48550/arxiv.2403.03407.

Thomas B Lawrence, Monika I Winn, and P Devereaux Jennings. The temporal dynamics of institutionalization. *Academy of management review*, 26(4):624–644, 2001.

Seth Lazar and Lorenzo Manuali. Can LLMs advance democratic values? *arXiv*, 2024.

Seth Lazar, Luke Thorburn, Tian Jin, and Luca Belli. The Moral Case for Using Language Model Agents for Recommendation. *arXiv*, 2024.

Margarita Leib, Nils C Köbis, Rainer Michael Rilke, Marloes Hagens, and Bernd Irlenbusch. The corruptive force of AI-generated advice. *arXiv*, 2021.

Joel Z Leibo, Alexander Sasha Vezhnevets, William A Cunningham, Sébastien Krier, Manfred Diaz, and Simon Osindero. Societal and technological progress as sewing an ever-growing, ever-changing, patchy, and polychrome quilt. *arXiv*, 2025. doi: 10.48550/arxiv.2505.05197.

Chao Li, Xing Su, Haoying Han, Cong Xue, Chunmo Zheng, and Chao Fan. Quantifying the Impact of Large Language Models on Collective Opinion Dynamics. *arXiv*, 2023. doi: 10.48550/arxiv.2308.03313.

Zhuoyan Li and Ming Yin. Utilizing Human Behavior Modeling to Manipulate Explanations in AI-Assisted Decision Making: The Good, the Bad, and the Scary. *arXiv*, 2024. doi: 10.48550/arxiv.2411.10461.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y Zou. Mapping the Increasing Use of LLMs in Scientific Papers. *arXiv*, 2024a.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024b.

Tao Lin, Kun Jin, Andrew Estornell, Xiaoying Zhang, Yiling Chen, and Yang Liu. User-Creator Feature Dynamics in Recommender Systems with Dual Influence. *arXiv*, 2024. doi: 10.48550/arxiv.2407.14094.

Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

Jinwei Lu, Yikuan Yan, Keman Huang, Ming Yin, and Fang Zhang. Do We Learn From Each Other: Understanding the Human-AI Co-Learning Process Embedded in Human-AI Collaboration. *Group Decision and Negotiation*, pp. 1–37, 2024. ISSN 0926-2644. doi: 10.1007/s10726-024-09912-x.

Adrian Lüders, Alejandro Dinkelberg, and Michael Quayle. Becoming "us" in digital spaces: How online users creatively and strategically exploit social media affordances to build up social identity. *Acta Psychologica*, 228:103643, 2022.

Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback Loop and Bias Amplification in Recommender Systems. *arXiv*, 2020. doi: 10.48550/arxiv.2007.13019.

S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14(1):4692, 2024. doi: 10.1038/s41598-024-53755-0.

Sandra G Mayson. Bias in, bias out. *YAle lJ*, 128:2218, 2018.

Lennart Meincke, Ethan R Mollick, and Christian Terwiesch. Prompting diverse ideas: Increasing ai idea variance. *arXiv preprint arXiv:2402.01727*, 2024.

Celestine Mendler-Dünner, Gabriele Carovano, and Moritz Hardt. An engine not a camera: Measuring performative power of online search. *arXiv preprint arXiv:2405.19073*, 2024a.

Celestine Mendler-Dünner, Gabriele Carovano, and Moritz Hardt. An engine not a camera: Measuring performative power of online search. *arXiv*, 2024b. doi: 10.48550/arxiv.2405.19073.

Carey K Morewedge, Sendhil Mullainathan, Haaya F Naushan, Cass R Sunstein, Jon Kleinberg, Manish Raghavan, and Jens O Ludwig. Human bias in algorithm design. *Nature Human Behaviour*, 7(11): 1822–1824, 2023.

Saumik Narayanan, Guanghui Yu, Chien-Ju Ho, and Ming Yin. How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making? *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 49–57, 2023. doi: 10.1145/3600211.3604709.

Christopher Nguyen. Value capture. *Journal of Ethics and Social Philosophy*, 27(3), 2024a. doi: 10.26556/jesp.v27i3.3048.

Christopher Nguyen. Value Capture. *Journal of Ethics and Social Philosophy*, 27(3), 2024b. doi: 10.26556/jesp.v27i3.3048.

Diana Bar-Or Nirman, Ariel Weizman, and Amos Azaria. Fool Me, Fool Me: User Attitudes Toward LLM Falsehoods. *arXiv*, 2024. doi: 10.48550/arxiv.2412.11625.

Aileen Oeberst and Roland Imhoff. Toward parsimony in bias research: A proposed common framework of belief-consistent information processing for a set of biases. *Perspectives on Psychological Science*, 18(6): 1464–1487, 2023.

Merrick R Osborne and Erica R Bailey. Me vs. the machine? subjective evaluations of human-and ai-generated advice. *Scientific Reports*, 15(1):3980, 2025.

Aviv Ovadya, Luke Thorburn, Kyle Redman, Flynn Devine, Smitha Milli, Manon Revel, Andrew Konya, and Atoosa Kasirzadeh. Toward Democracy Levels for AI. *arXiv*, 2024. doi: 10.48550/arxiv.2411.09222.

Vishakh Padmakumar and He He. Does Writing with Language Models Reduce Content Diversity? *arXiv*, 2023.

Scott E Page et al. Path dependence. *Quarterly Journal of Political Science*, 1(1):87–115, 2006.

Cezara Panait and Cameran Ashraf. Ai algorithms–(re) shaping public opinions through interfering with access to information in the online environment. *Europuls Policy Journal*, 1(1):46–64, 2021.

Luca Pappalardo, Emanuele Ferragina, Salvatore Citraro, Giuliano Cornacchia, Mirco Nanni, Giulio Rossetti, Gizem Gezici, Fosca Giannotti, Margherita Lalli, Daniele Gambetta, Giovanni Mauro, Virginia Morini, Valentina Pansanella, and Dino Pedreschi. A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions. *arXiv*, 2024.

Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5(10):1076–1086, 2023. doi: 10.1038/s42256-023-00720-7.

Dino Pedreschi, Luca Pappalardo, Emanuele Ferragina, Ricardo Baeza-Yates, Albert-László Barabási, Frank Dignum, Virginia Dignum, Tina Eliassi-Rad, Fosca Giannotti, János Kertész, et al. Human-ai coevolution. *Artificial Intelligence*, 339:104244, 2025.

Nicola Perra and Luis E. C. Rocha. Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific Reports*, 9(1):7261, 2019. doi: 10.1038/s41598-019-43830-2.

Andrew J Peterson. AI and the Problem of Knowledge Collapse. *arXiv*, 2024.

Buu Phan, Marton Havasi, Matthew Muckley, and Karen Ullrich. Understanding and mitigating tokenization bias in language models. *arXiv preprint arXiv:2406.16829*, 2024.

Jason Phang, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, et al. Investigating affective use and emotional well-being on chatgpt. *arXiv preprint arXiv:2504.03888*, 2025.

Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Fernando P Santos, Yong Li, and James Evans. Emergence of human-like polarization among large language model agents. *arXiv*, 2025. doi: 10.48550/arxiv.2501.05171.

Tiziano Piccardi, Martin Saveski, Chenyan Jia, Jeffrey T Hancock, Jeanne L Tsai, and Michael Bernstein. Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity. *arXiv*, 2024.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. *arXiv*, 2024.

Tianyi Alex Qiu, Zhonghao He, Tejasveer Chugh, and Max Kleiman-Weiner. The lock-in hypothesis: Stagnation by algorithm. *ICML 2025*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Moritz Reis, Florian Reis, and Wilfried Kunde. Influence of believed AI involvement on the perception of digital medical advice. *Nature Medicine*, 30(11):3098–3100, 2024. ISSN 1078-8956. doi: 10.1038/s41591-024-03180-7.

Yi Ren, Shangmin Guo, Linlu Qiu, Bailin Wang, and Danica J Sutherland. Language Model Evolution: An Iterated Learning Perspective. *arXiv*, 2024. doi: 10.48550/arxiv.2404.04286.

Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

Evan F Risko and Sam J Gilbert. Cognitive offloading. *Trends in cognitive sciences*, 20(9):676–688, 2016.

Michael J Ryan, William Held, and Diyi Yang. Unintended Impacts of LLM Alignment on Global Representation. *arXiv*, 2024.

Abel Salinas, Louis Penafiel, Robert McCormack, and Fred Morstatter. " im not racist but...": Discovering bias in the internal knowledge of large language models. *arXiv preprint arXiv:2310.08780*, 2023.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

Nick Schuster and Seth Lazar. Attention, moral skill, and algorithmic recommendation. *Philosophical Studies*, 182(1):159–184, 2025. ISSN 0031-8116. doi: 10.1007/s11098-023-02083-6.

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2024. doi: 10.1145/3613904.3642459.

Minkyu Shin, Jin Kim, Bas Van Opheusden, and Thomas L Griffiths. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences*, 120(12):e2214840120, 2023.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *arXiv*, 2024.

Waldomiro J Silva Filho, Maria Virginia M Dazzani, Luca Tateo, Rodrigo Gottschalk Sukerman Barreto, and Giuseppina Marsico. He knows, she doesn't? epistemic inequality in a developmental perspective. *Review of General Psychology*, 27(3):231–244, 2023.

Felix M Simon and Luisa Fernanda Isaza-Ibarra. Ai in the news: reshaping the information ecosystem? 2023.

Aaditya K Singh and DJ Strouse. Tokenization counts: the impact of tokenization on arithmetic in frontier llms. *arXiv preprint arXiv:2402.14903*, 2024.

Betsy Sparrow, Jenny Liu, and Daniel M Wegner. Google effects on memory: Cognitive consequences of having information at our fingertips. *science*, 333(6043):776–778, 2011.

Jessica Su, Aneesh Sharma, and Sharad Goel. The effect of recommendations on network structure. In *Proceedings of the 25th international conference on World Wide Web*, pp. 1157–1167, 2016.

Allen Daniel Sunny. Preliminary quantitative study on explainability and trust in ai systems. *arXiv preprint arXiv:2510.15769*, 2025.

Nicole Miu Takagi. Banning of chatgpt from educational spaces: A reddit perspective. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pp. 179–194, 2023.

Rohan Taori and Tatsunori B Hashimoto. Data Feedback Loops: Model-driven Amplification of Dataset Biases. *arXiv*, 2022.

Irmgard Teschke, Claudia AF Wascher, Madeleine F Scriba, Auguste MP von Bayern, V Huml, B Siemers, and Sabine Tebbich. Did tool-use evolve with enhanced physical cognitive abilities? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1630):20120418, 2013.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024. ISSN 0036-8075. doi: 10.1126/science.adq2852.

Brian Thompson, Mehak Preet Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism. *arXiv*, 2024.

ELINA TREYGER, JOSEPH MATVEYENKO, and LYNSAY AYER. Manipulating minds. 2025.

A. Valyaeva. Ai has already created as many images as photographers have taken in 150 years. *Everypixel Journal*, 2024. URL `https://journal.everypixel.com/ai-image-statistics`.

Lucía Vicente and Helena Matute. Humans inherit artificial intelligence biases. *Scientific Reports*, 13(1): 15737, 2023.

Christian Wagner and Ling Jiang. Death by AI: Will large language models diminish Wikipedia? *Journal of the Association for Information Science and Technology*, 2025. ISSN 2330-1635. doi: 10.1002/asi.24975.

Basil Wahn, Laura Schmitz, Frauke Nora Gerster, and Matthias Weiss. Offloading under cognitive load: Humans are willing to offload parts of an attentionally demanding task to an algorithm. *Plos one*, 18(5): e0286102, 2023.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pp. 1–12, 2025.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. Decoding Echo Chambers: LLM-Powered Simulations Revealing Polarization in Social Networks. *arXiv*, 2024. doi: 10.48550/arxiv.2409.19338.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE transactions on knowledge and data engineering*, 29(12):2724–2743, 2017.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. *arXiv*, 2025. doi: 10.48550/arxiv.2501.13381.

Joshua P White, Carter Allen, Lucius Caviola, Thomas H Costello, and David G Rand. Increasing the effectiveness of charitable giving using human-ai dialogues. *preprint*, 2024.

John Wihbey. AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge? *SSRN Electronic Journal*, 2024. doi: 10.2139/ssrn.4805026.

Marty J Wolf, K Miller, and Frances S Grodzinsky. Why we should have seen that coming: comments on microsoft's tay" experiment," and wider implications. *Acm Sigcas Computers and Society*, 47(3):54–64, 2017.

Fan Wu, Emily Black, and Varun Chandrasekaran. Generative Monoculture in Large Language Models. *arXiv*, 2024.

Weiqi Xu and Fan Ouyang. A systematic review of ai role in the educational system based on a proposed conceptual framework. *Education and Information Technologies*, 27(3):4195–4223, 2022.

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, and Iyad Rahwan. Empirical evidence of large language model's influence on human spoken communication. *arXiv*, 2024. doi: 10.48550/arxiv.2409.01754.

Nicolas Yax, Hernan Anlló, and Stefano Palminteri. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51, 2024.

Doron Yeverechyahu, Raveesh Mayya, and Gal Oestreicher-Singer. The impact of large language models on open-source innovation: Evidence from github copilot. *arXiv preprint arXiv:2409.08379*, 2024.

Chunpeng Zhai, Santoso Wibowo, and Lily D Li. The effects of over-reliance on ai dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1):28, 2024.

Zhaowei Zhang, Fengshuo Bai, Mingzhi Wang, Haoyang Ye, Chengdong Ma, and Yaodong Yang. Incentive compatibility for ai alignment in sociotechnical systems: Positions and prospects. *arXiv preprint arXiv:2402.12907*, 2024.

Zhaowei Zhang, Minghua Yi, Mengmeng Wang, Fengshuo Bai, Zilong Zheng, Yipeng Kang, and Yaodong Yang. EuroCon: Benchmarking parliament deliberation for political consensus finding. *arXiv*, 2025. doi: 10.48550/arxiv.2505.19558.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, 2024.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.